一种自下而上的人脸检测算法

张　宁，伍萍辉†

(河北工业大学 电子信息工程学院 天津市电子材料与器件重点实验室，天津 300401)

摘　要：针对在非控条件下的人脸检测经常遇到的问题，如复杂的人脸姿态表情、严重的人脸遮挡、外界环境背景复杂、光照条件差、小人脸等提出了一种自下而上的人脸检测方法。自下而上的人脸检测是基于深度学习的，先进行人脸相关关键点检测和关键点之间的位置关系检测再进行人脸检测。网络结构采用稠密网络进行图像特征提取，提取到的特征传送给 6 个级联网络，每个级联网络由两个分支网络构成，分支网络 1 用来预测人脸相关关键点位置坐标，分支网络 2 用来预测关键点之间的位置关系。利用得到的关键点位置和位置关系进行人脸检测。在 FDDB 测试集上进行了验证，取得了 0.98 的成绩，并可以在输入图像分辨率为 1920×1080 的情况下，能检测到的最小人脸分辨率为 10×10，使用 GPU Nvidia Gefore GTX 1070 最快能达到 17 fps。

关键词：人脸检测；深度学习；关键点检测；自下而上

中图分类号：TP391.41　　doi: 10.3969/j.issn.1001-3695.2018.01.0032

Bottom-up face detection algorithm

Zhang Ning, Wu Pinghui†

(*Tianjin Key Laboratory of Electronic Materials & Devices*, *School of Electronic & Information Engineering Hebei University of Technology*, *Tianjin* 300401, *China*)

Abstract: Faced with the problems often encountered in face detection under non-controlled conditions, such as complex facial expression, serious face occlusion, complex external environment, poor lighting conditions, tiny face, etc. A bottom-up face detection method is proposed. Bottom-up face detection is based on deep learning, face detection and key points of the first position-related key detection and then face detection. Convolution neural network structure using dense network for image feature extraction, the extracted features are transmitted to 6 cascaded networks, each of which consists of two branch networks. Branch network 1 is used to predict the coordinates of face-related key points. Branch network 2 is used to predict the position between key points relationship. Face detection is performed by using the obtained key point position and position relationship. The FDDB test set was verified and achieved 0.98 results, and the smallest face resolution 10×10 can be detected at the resolution of input image 1920×1080, used the GPU Nvidia Geforce GTX 1070 for up to 17fps video detection.

Key words: face detection; deep learning; key point detection; bottom-up

0　引言

在“911”事件之后,人脸检测识别逐渐成为国际反恐和安全防范重要的手段之一。随着人工智能时代的到来，机器视觉领域得到了进一步的发展。目前在目标检测、目标识别、图像分割等机器视觉项目大赛中，那些基于深度学习的算法准确性明显高于传统的手工设计特征提取的算法准确性。尽管基于深度学习的算法可以解决大部分目标检测的问题，但是目前仍有小目标检测的问题深度学习解决的还不是很好。为解决人脸检测中小脸的问题，提出了一种自下而上的人脸检测方法。小脸就是指在图像分辨率为1920×1080情况下，人脸成像分辨率不大于 20×20。根据国内外对人脸检测算法的研究热度来看，检测小脸仍然是值得挑战的项目。

国外主要有 MIT、CMU、CORNELL 等大学和 Google DeepMind、Facebook OpenMind 等一些研究机构，国内主要有清华大学、北京大学、中国科学院计算机所、中国科学院自动化所等一系列高校和研究机构都致力于解决现实生活中的人脸检测定位问题。对于人脸检测的挑战也主要有三方面：a)人脸姿态、表情、遮挡等内在的因素；b)受外界环境因素干扰，如光照不均匀或者光照条件很差，人脸成像带有动态模糊等外在因素；c)时效性，若是算法过于复杂，计算量很大会直接影响算法时效性，无法满足基于视频监控的人脸实时性检测要求。

收稿日期：2018-01-18；**修回日期：**2018-03-15

作者简介：张宁（1992-），男，硕士研究生，主要研究方向为机器视觉、物体检测；伍萍辉（1970-），女（通信作者），教授，博士，主要研究方向为电子电路设计、微机测控技术、智能控制（wuphui@126.com）.

针对以上三个方面的挑战，提出了一种自下而上的人脸检测方法。通过设计卷积神经网络结构，构造目标函数，再利用反向传播算法来不断更新模型中的参数，直到目标函数最小化。经过在 FDDB[1]人脸数据集上的测试，得到了很好的结果，同时也在监控视频的场景下进行了测试，可以满足在输入图像分辨率为1920×1080的监控场景下，能检测到的最小人脸分辨率为10×10，使用 GPU Nvidia Gefore GTX 1070 最快能达到 17 fps。

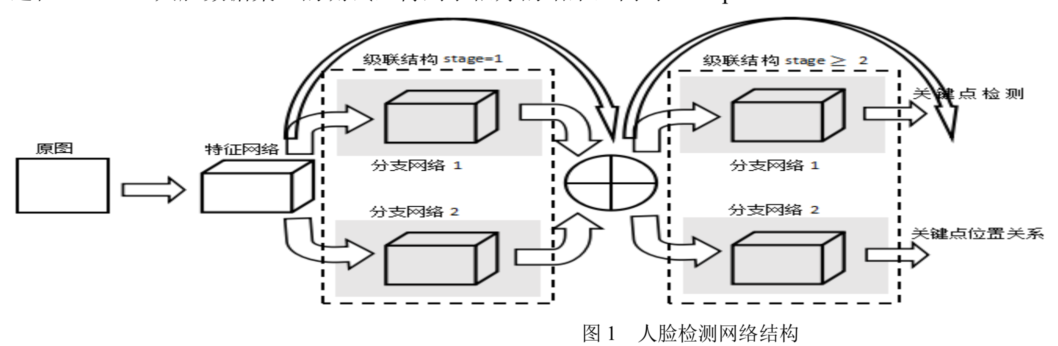


图 1 人脸检测网络结构

1 人脸检测算法概述

当前基于深度学习的人脸检测算法性能普遍优于传统的手工设计特征提取的人脸检测算法[2]。主流的人脸检测算法有采用级联结构的 Cascade CNN[3]、MTCNN[4]、ICC[5]，采用端到端的 Finding Tiny Faces[6]，还有根据基于卷积神经网络的目标检测[7]改成的人脸检测 [8]。下面简单地介绍其中几种算法。Cascade CNN 采用 6 个级联的浅层网络，其中 3 个进行人脸/非人脸的二分类，另外 3 个进行人脸框的校准，也就是进行 45 分类，利用标定人脸框左上角的坐标值（x,y）和标定人脸框的宽度 w 和高度 h，n 取 1～45 的整数，采用式（1）～（4）进行人脸框的校准。该算在标准 VGA 图像基于 CPU 的检测中可以达到 14 fps，基于 GPU 可以达到 100 fps。

$$x_n \in \{-0.17,0,0.17\} \quad (1)$$

$$x-\frac{x_n w}{s_n},y-\frac{y_n h}{s_n},\frac{w}{s_n},\frac{h}{s_n} \quad (2)$$

$$s_n \in \{0.83,0.91,1.0,1.1,10.21\} \quad (3)$$

$$y_n \in \{-0.17,0,0.17\} \quad (4)$$

MTCNN 相对于 Cascade CNN，也采用级联方式进行人脸检测，但是不同之处是 MTCNN 采用三层级联网络结构，三个网络同时进行训练人脸和非人脸二分类、人脸候选框回归、人脸对齐操作，属于多任务级联网络。Hu 等人[6]提出的 Finding Tiny Faces 针对解决人脸检测中的小脸问题得到了很好的效果，该方法提出可以充分利用人脸上下文信息和缩放图像分辨率进行小脸检测，是一个端到端的网络结构。Jiang 等人[8]将 Faster R-CNN[9]进行目标检测的网络利用迁移学习改成了人脸检测网络。本文提出的一种自下而上的人脸检测算法将与以上算法在 FDDB 测试集上进行对比实验。

2 自下而上的人脸检测

2.1 卷积神经网络原理

目前卷积神经网络在图像领域和语音领域取得了很好的成绩。卷积神经网络是一个层级结构，由卷积层、激活层[10]、池化层、归一化层[11]、全连接层等构成，最后还要连接损失层。卷积神经网络的计算分为两个过程，即前向传播和反向传播[12]。前向传播指的就是输入的数据经过卷积神经网络的层级结构最终会到达损失层，对每一层的作用进行一个简要说明，卷积层进行卷积操作提取图像特征信息，激活层引入非线性特性，池化层进行降维操作，可以减少整个网络计算时间，归一化层可以加快网络的收敛速度，使得到的模型具有一定的泛化能力，全连接层起到一个类似“分类器”的作用。反向传播可以简单地理解为复合函数的链式法则，利用经验风险最小化，将损失函数的计算值逼近为 0，这样预测值和真实值就可以相差最小或者一样了。为了达到这一目的，必须要更新各个层级结构中的权重值 w 和偏置 b，只有 w，b 改变了才能改变预测值，而 w 和 b 的更新完全依靠求导的链式法则。

2.2 网络结构

一种自下而上的人脸检测算法是受 Cao 等人[13]的启发。自下而上指的是通过先检测人脸相关关键点和关键点之间的位置关系再确定人脸大小及位置。人脸相关关键点检测的同时还进行了关键点之间的位置关系检测，这样做的好处是可以更加准确地检测出人脸位置，在很大程度上克服了人脸表情、人脸姿态、人脸遮挡以及背景环境复杂等带来的影响。而且受 Finding Tiny Face 和 ICC 启发，结合上下文信息，利用了肩膀和脖子这些和人脸有关联的部位，能检测到的最小人脸分辨率为10×10，又通过优化网络结构，在输入图像分辨率为1920×1080的情况下，使用 GPU Nvidia Gefore GTX1070 最快能达到 17 fps。

整个网络结构如图 1 所示。采用分支网络 1 来预测人脸相关关键点位置,分支网络 2 来预测关键点位置之间的关系。特征网络采用稠密网络 (DenseNet) [14]意味着每个卷积层之间都直接连接着。普通的网络层是第 L 层有 L 个连接,如图 2（a）所示,第

4 个卷积层有 4 个连接,一个连接是卷积层 4 和卷积层 3 之间的连接,剩下的 3 个连接是卷积层 4 之前的层之间的连接。稠密网络第 L 层有 $\frac{L(L+1)}{2}$ 个连接,如图 2（b）所示,第 4 个卷积层有 10 个连接。每一个卷积层都与之前的

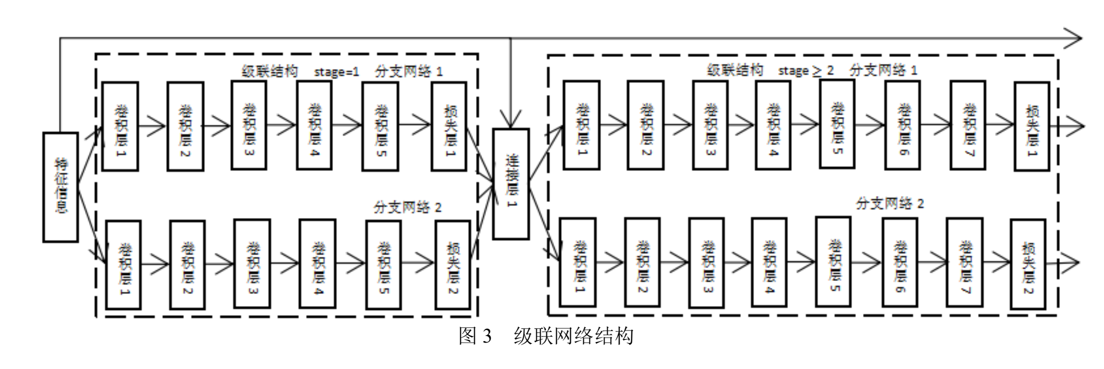
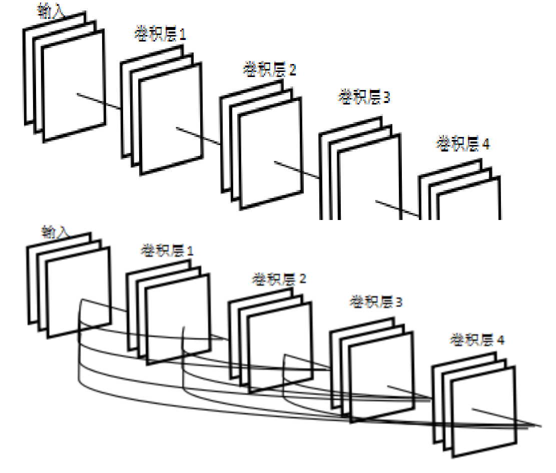


图 3 级联网络结构



（b）4 层卷积网络的稠密网络

图 2 稠密网络与普通网络对比

所有卷积层直接连接,这样做有三个好处：a)避免了反传时的梯度消失问题；b)低层的语义信息可以直接传递给更高层,使特征可以重复利用；c)可以减少超参数,降低了过拟合的风险,这一点满足了网络加深的同时既提升精度也保证了速度。

对图 1 网络结构的级联结构进行展开就如图 3 所示,可以看到级联结构处于第一阶段也就是 stage=1 时,是第一级级联,采用全卷积结构,没有使用全连接层。第一级级联中分支网络 1 和 2 一样都有 5 个卷积层,参考 AlexNet 网络结构,并在卷积层 1、2、3、4 后面紧跟激活层 Relu,后面的卷积层 5 参考 FCN[15]网络结构,采用的卷积核为1×1,代替了全连接层的作用,其余卷积层采用的卷积核为3×3,最后面连接的是损失层。stage≥2 时,分支网络 1 和 2 都采用 7 个卷积层,卷积层 1 到 6 卷积核大小为3×3,卷积层 7 卷积核大小为1×1,最后连接一个损失层,损失层 1 和 2 分别采用损失函数如式（5）（6）所示。

$$L=\sum_{i=1}^{n}\sum_{j=1}^{m}\|\hat{y}_j^i-y_j^i\|_2^2 \quad (5)$$

其中：n 表示关键点的数量，m 表示人脸的数量；$\hat{y}_j^i$ 表示第 j 个人脸的第 i 个人脸关键点的真实值；y_j^i 表示第 j 个人脸的第 i 个人脸关键点的预测值。

$$L=\sum_{k=1}^{l}\sum_{j=1}^{m}\|\hat{y}_j^k-y_j^k\|_2^2 \quad (6)$$

其中：l 表示关键点位置关系的数量；m 表示人脸的数量；$\hat{y}_j^k$ 表示第 j 个人脸的第 k 个人脸相关关键点位置关系的真实值；y_j^k 表示第 j 个人脸的第 k 个人脸相关关键点位置关系的预测值。

分支网络 1 进行人体 8 个关键点预测（右耳、右眼、鼻子、左眼、左耳、脖子、右肩、左肩）,分支网络 2 预测相邻两个关键点之间的位置关系,如图 4 所示。值得注意的是,stage≥2 时输入不仅来自前一阶段的特征信息,还有来自 DenseNet 网络提取的特征信息。

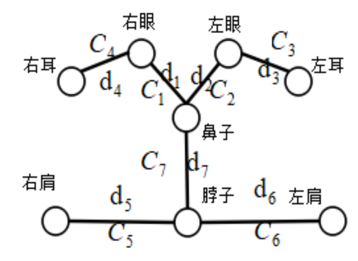


图 4 检测的关键点

C_i 为相邻两个关键点之间位置关系的置信度得分，d_i 为相邻两个关键点之间的距离，i 取 1～7 个整数。

d_i 根据距离计算式（7），x_a、y_a 分别为相邻两个关键点其中一个点的横坐标和纵坐标，x_b、y_b 分别为另一个点的横坐标和纵坐标。

$$d_i=\sqrt{(x_a-x_b)^2+(y_a-y_b)^2} \quad (7)$$

2.3 人体检测流程

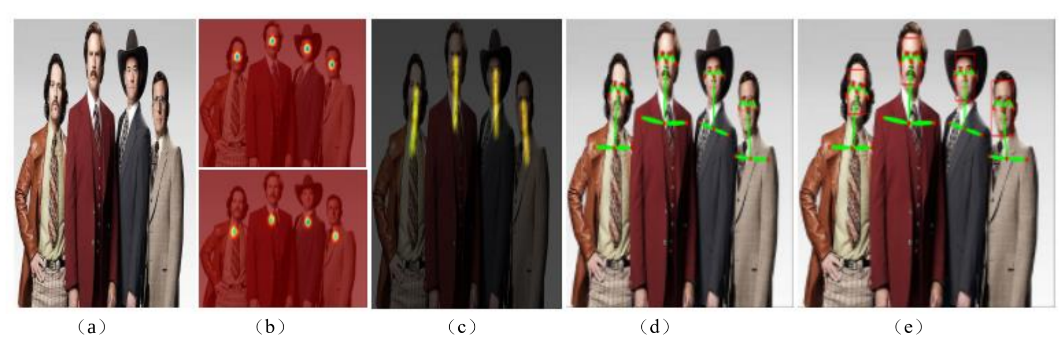
整体的人脸检测流程如图 4 所示。其中图 4（a）为输入图像；（b）是分支网络 1 关键点检测可视化结果；（c）分支网络 2 用来进行关键点位置关系预测可视化结果；网络输出的检

测结果如图 4（d）所示；利用人脸关键点位置之间的逻辑关系就可以画出人脸检测框，如图 4（e）所示。

人脸框的置信度得分 C_{face} 根据式（8）

$$C_{\text{face}} = \sum_{i=1,2,3,4,7} \frac{d_i}{\sum d_i} \cdot C_i \qquad (8)$$

计算人脸框的算法流程如下：



（a）　（b）　（c）　（d）　（e）

图 4　人脸检测流程

a)根据网络模型可获得关键点的位置坐标以及相邻两点间位置关系的置信度得分，利用关键点鼻子坐标（$x_{鼻子}$，$y_{鼻子}$）。

b)利用式（7）计算 d_1, d_2, d_3, d_4。

c)计算人脸框左上角坐标位置（x_{l_up}，y_{l_up}），右下角坐标位置（x_{r_down}，y_{r_down}），计算公式如（8）～所示。人脸在水平方向分为右侧脸（$d_1 > d_2$）,正脸（$d_1 = d_2 \neq 0$）,左侧脸（$d_1 < d_2$）三种姿态。

$$x_{l_\text{up}} = \begin{cases} x_{鼻子} - d_4 - d_1, & d_1 > d_2 \\ x_{鼻子} - 1.2 \cdot d_1, & d_1 = d_2 \neq 0 \\ x_{鼻子} - d_1, & d_1 < d_2 \end{cases} \qquad (8)$$

$$y_{l_\text{up}} = \begin{cases} y_{鼻子} - 1.1 \cdot d_1 - d_2, & d_1 > d_2 \\ y_{鼻子} - 1.6 \cdot d_1, & d_1 = d_2 \neq 0 \\ y_{鼻子} - 1.1 \cdot d_2, - d_1 & d_1 < d_2 \end{cases} \qquad (9)$$

$$x_{\text{r}_down} = \begin{cases} x_{鼻子} + d_2, & d_1 > d_2 \\ x_{鼻子} + 1.2 \cdot d_1, & d_1 = d_2 \neq 0 \\ x_{鼻子} + d_2 + d_3, & d_1 < d_2 \end{cases} \qquad (10)$$

$$y_{\text{r}_down} = \begin{cases} y_{鼻子} + 1.1 \cdot d_2 + d_1, & d_1 > d_2 \\ y_{鼻子} + 1.6 \cdot d_1, & d_1 = d_2 \neq 0 \\ y_{鼻子} + 1.1 \cdot d_1 + d_2, & d_1 < d_2 \end{cases} \qquad (11)$$

d)利用 NMS 将重叠的人脸框进行去除，剩下的人脸框映射到原图，得到人脸检测的最终结果。

3　训练过程

3.1　训练数据准备

训练数据采用微软提供的 COCO[16]数据集，使用 caffe 进行模型训练。该数据集的训练集包含超过了 10 万个人的标注数据，约有 1 百万个人体关键点，但本算法只利用部分人体关键点（如人的左右耳朵、左右眼睛、鼻子、左右肩膀和脖子）。同时在训练过程中采用数据增广的策略，对训练样本进行旋转、翻转等，目的是增加模型的泛化能力。部分训练样本的数据如图 5 所示，可视化了关键点和关键点位置关系，同时还利用掩膜遮盖了未进行标注的人体关键部位，这将有助于网络模型学习人体关键点，其他未标注的人体关键点不会当作负样本影响网络学习。

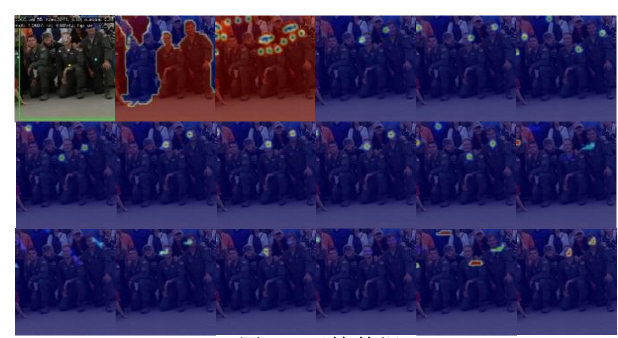


图 5　训练数据

3.2　端到端训练

训练阶段采用 6 个级联网络（stage=6），其中从第二个级联网络到第六个级联网络的结构都是一样的，相当于将每个级联网络结构模块化。每个级联网络结构都有损失层，防止网络过深梯度无法反传。整个训练过程是端到端的，并不需要逐个训练每个级联网络。初始学习率设置为 0.005，迭代 600 000 次。

4　实验结果

FDDB 数据集有 2 845 张图像、5 171 个人脸。收集的这些人脸包含各种姿态、各种面部表情、不同光照环境、不同背景环境、不同分辨率以及不同聚焦环境下，有灰度图也有彩色图。

在标注中，人脸的宽度和高度均不小于 20 个像素点，同时采用椭圆标注，可以更进一步地贴合人脸。本文提出的自下而上的人脸检测算法在 FDDB 数据集上的检测效果如图 6 所示。其中展示了在光照条件差、人脸被遮挡、人脸表情姿态很复杂以及人脸模糊情况下的检测效果。FDDB ROC 曲线如图 7 所示。

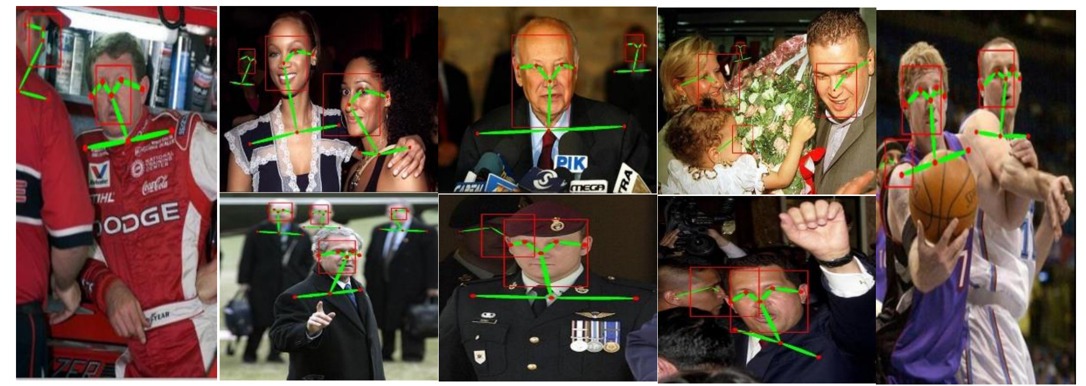
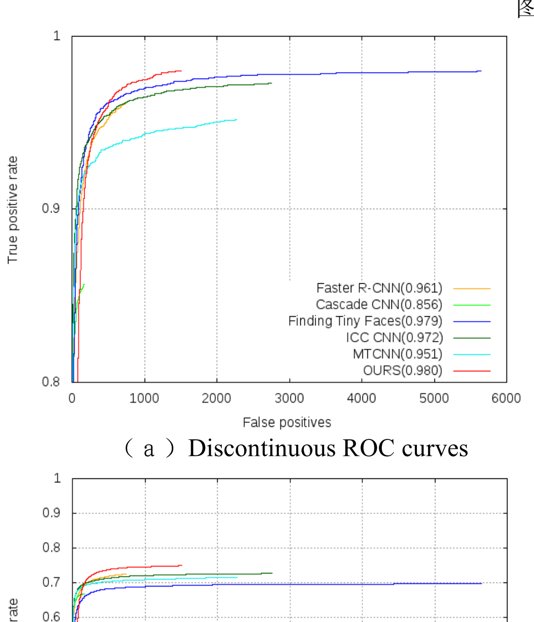


图 6　FDDB 数据集检测结果

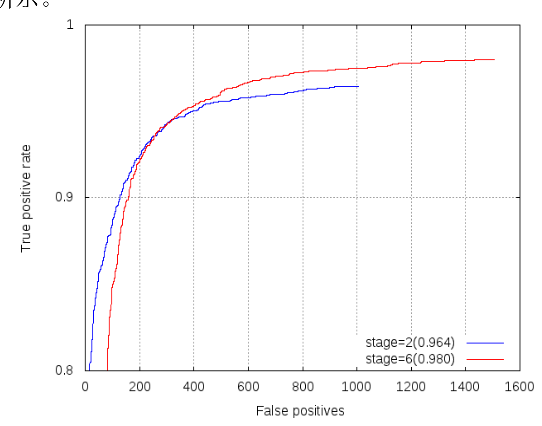


（a）Discontinous ROC curves

（b）Continous ROC curves

图 7　FDDB ROC　曲线

为了能将本文算法实用化，就需要对非控条件下的人脸做到实时检测。本文算法的网络结构是比较深的，尽管使用 DenseNet 减少了滤波器个数，但模型还是比较大的，参数还是比较多的。优化网络结构成为了重中之重。在测试阶段，因为后面的级联网络结构是模块化的，而且每个模块都可以独立的完成网络前传，这样就可以由 6 个级联模块减少到最少 2 个级联模块。通过在 FDDB 数据集上测试，发现精度只减少了 0.016，如图 8（a）所示，但是速度却提升了将近一倍。整个对比实验的结果如表 1 所示。采用的处理器为英特尔 core 7-7700HQ@2.80 GHz 八核，内存 RAM16 GB，显存 8 GB，显卡 Nvidia Gefore GTX 1070，输入检测视频的图像分辨率为 1920×1080，检测到的最小脸分辨率为 10×10。在非控场景下的人脸检测效果图如图 9 所示。



（a）Discontinous ROC curves

（b）Continous ROC curves

图 8　stage=2 和 stage=6 FDDB ROC　曲线

表 1　算法运行时间对比

级联结构（stage）	平均检测帧速率（fps）	ROC 曲线（FDDB）
2	16	0.964
6	8	0.980

为了比较 MTCNN、Finding Tiny Faces 和本文算法（OURS），分别在光线较暗、姿态表情复杂、图像分辨率低、存在人脸遮

挡四种情况下进行了实验。实验结果如图 10 所示。其中第一行是 MTCNN 在以上四种情况下的检测结果，不难发现 MTCNN 存在很高的误检率，而且很难检测到光线较暗的人脸和分辨率较低的侧脸；第二行是 Finding Tiny Faces



图 9　非控场景下人脸检测

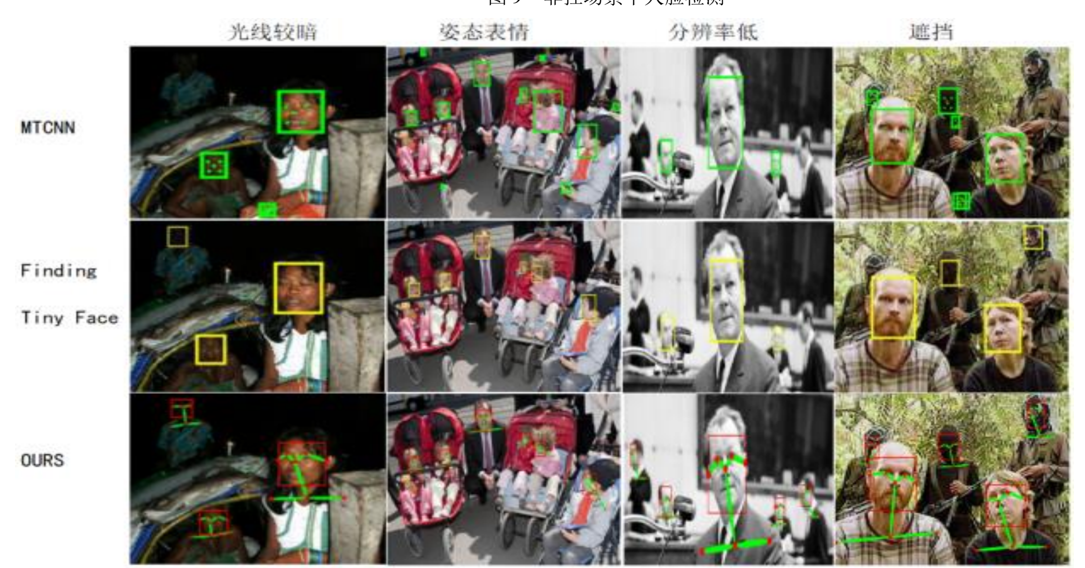


图 10　三种算法对比

在以上四种情况下的检测结果；第三行是 OURS 在以上四种情况下的检测结果，可以看出 Finding Tiny Faces 和 OURS 检测结果相差不大，只是在分辨率较低的情况下，OURS 对侧脸检测比 Finding Tiny Faces 要好一些。本文还进行了三种算法检测时间的对比实验，实验条件如下：使用 GPU Nvidia Gefore GTX 1070 对来自 AFLW[17]和 FDDB 的数据集随机取 3 000 张，这 3 000 张图片中有的图片只包含一个人脸，有的图片包含多个人脸。三种算法对每个图片检测 100 次取平均检测时间，结果如图 11 所示。图中横坐标为人脸数量，纵坐标为检测时间（单位为 ms）。可以看到 Finding Tiny Faces 检测时间最长且与人脸个数成正比。MTCNN 检测时间与人脸个数成正比，随着人脸个数的增多，检测时间已经越来越高于 OURS 检测时间了。OURS 受到人脸个数的影响相对较小，但也呈现出随着人脸个数增多检测时间变长的趋势。

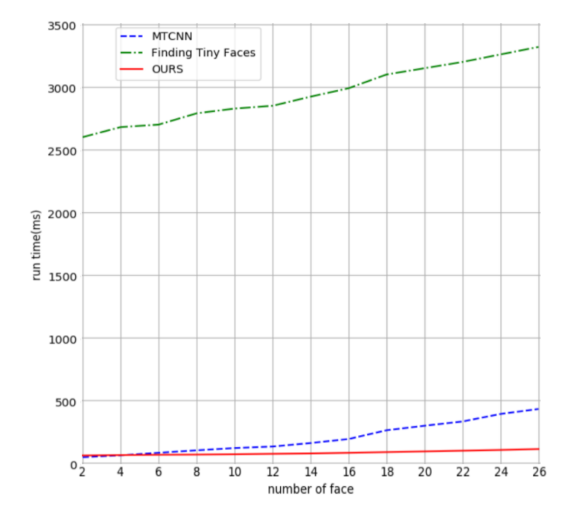


图 11　人脸数量和检测时间关系

5 结束语

本文提出的自下而上的人脸检测算法，先通过关键点检测和关键点位置关系检测，再进行人脸检测，得到了很好的检测效果。在FDDB数据集上与其他几种基于深度学习的人脸检测算法对比得到了比较高的准确率。更重要的是，本文提出的算法在面对复杂的人脸表情、人脸姿态、人脸遮挡等内在因素干扰，外界背景环境复杂、光照条件差等诸多外界干扰因素下都能得到不错的人脸检测效果，且在检测时间上受到人脸个数的影响较小。经实测监控场景下的视频最快时能达到17 fps，使用GPU Nvidia Gefore GTX 1070在输入图像分辨率1920×1080能检测到最小脸为10×10。

参考文献：

[1] Jain V, Learned-Miller E. FDDB: a benchmark for face detection in unconstrained settings, UMass Amherst Technical Report [R]. 2010.

[2] 田原嫄，姚萌萌，潘敏凯，等. 基于YCbCr肤色检测与AdaBoost联级算法的嘴部特征定位 [J]. 计算机应用研究, 2017, 34 (3): 933-935. (Tian Yuanyuan, Yao Mengmeng, Pan Minkai, *et al.* Human' smouth location based on YCbCr complexion detection and AdaBoost cascade connection method [J]. Application Research of Computers, 2017, 34 (3): 933-935.)

[3] Li Haoxiang, Lin Zhe, Shen Xiaohui, *et al.* A convolutional neural network cascade for face detection [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: [s. n.] , 2015: 5325-5334.

[4] Zhang Kaipeng, Zhang Zhanpeng, Li Zhifeng, *et al.* Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Processing Letters, 2016, 23 (10): 1499-1503.

[5] Zhang Kaipeng, Zhang Zhanpeng, Wang Hao, *et al.* Detecting faces using inside cascaded contextual CNN [C]// Proc of IEEE International Conference on Computer Vision. Venice, Italy: [s. n.] , 2017: 3171-3179.

[6] Hu Peiyun, Ramanan D. Finding tiny faces [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii: IEEE Computer Society. 2017: 1522-1530.

[7] 李旭冬，叶茂，李涛. 基于卷积神经网络的目标检测研究综述 [J]. 计算机应用研究, 2017, 34 (10): 2881-2886, 2891. (Li Xudong, Ye Mao, Li Tao. Review of object detection based on convolutional neural networks [J]. Application Research of Computers, 2017, 34 (10): 2881-2886, 2891.)

[8] Jiang Huaizu, Learned-Miller Erik. Face detection with the faster R-CNN [C]// Proc of IEEE International Conference on Automatic Face & Gesture Recognition. Washington, DC: [s. n.] , 2017: 650-657.

[9] Ren Shaoqing, He Kaiming, Girshick Ross, *et al.* Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Trans on Pattern Anal Mach Intell, 2015, 39 (6): 1137-1149.

[10] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: [s. n.] , 2012: 1097-1105.

[11] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]// Proc of the 32nd International Conference on Machine Learning. PMLR, Lille, France: [s. n.] , 2015: 448-456.

[12] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors [J]. Readings in Cognitive Science, 1988, 323 (6088): 399-421.

[13] Cao Zhe, Simon T, Wei Shihen, *et al.* Realtime multi-person 2D pose estimation using part affinity fields [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu. Hawaii, USA: [s. n.] , 2017: 1302-1310.

[14] Huang Gao, Liu Zhuang, Weinberger K Q. Densely connected convolutional networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, United State: IEEE Computer Society, 2016: 3276-3281.

[15] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2017, 39 (4): 640.

[16] Lin Tsungyi, Maire M, Belongie S, *et al.* Microsoft COCO: common objects in context [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2014: 740-755.

[17] Köstinger M, Wohlhart P, Roth P M, et al. Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization [C]// Proc of IEEE International Conference on Computer Vision Workshops. Barcelona, Spain: [s. n.] , 2011: 2144-2151.